Подбор гиперпараметров и метод аугментации данных для различных опорных моделей Mask R-CNN

А. Д. Егоров[1], М. С. Резник[2]
Национальный исследовательский ядерный университет МИФИ
[1]ADEgorov@mephi.ru, [2]maximrez@mail.ru

***Аннотация*. Одна из наиболее сложных задач компьютерного зрения – задача определения действия объекта. Для решения этой задачи необходимо знать информацию о положении ключевых точек объекта конкретного типа. Сложность задачи определяется так же тем, что для обучения моделей, которые могут распознать ключевые точки, требуется большой объём сложно-организованных данных. В рамках данной работы решается задача определения положения ключевых точек биологического объекта. Эта информация необходима для классификации действий объекта и принятия технической системой управленческого решения о взаимодействии с объектом. В связи с недостатком данных для обучения, приводится метод для получения дополнительных данных для обучения (аугментация данных), а также проводится тестирование различных семейств опорных моделей в рамках сетей типа R-CNN на по-разному аугментированных данных, с различными вариантами оптимизаторов, скоростью обучения, количеством эпох обучения и батчей. Достигнутая точность на тестовой выборке: больше 90 %. Использование опорных моделей семейства ResNet позволило добиться большей точности работы, которая составила более 93%, тогда как использование опорных моделей из семейства MobileNet при точности порядка 90 % позволило добиться скорости обработки каждого кадра в 3 раза большей (в среднем), чем при использовании опорных моделей семейства ResNet.**

Ключевые слова: Keypoint R-CNN, опорные модели, классификация скелета, аугментация данных, подбор гиперпараметров

I. Общая постановка задачи и проблематизация

Развитие технологии компьютерного зрения и увеличение доступных вычислительных мощностей привело к тому, что был сильно расширен круг задач, которые можно решать с помощью данной технологии. Важной задачей компьютерного зрения на текущий момент является описание действия объекта по изображению или видеопотоку, в котором он обнаружен [1, 2]. В технических системах подобная информация используется для выработки управленческого решения во время взаимодействия с объектом в условиях неопределённости. При этом, в системах зачастую используются сложные обучаемые модели для достижения высокого качества работы средств управления. Для решения таких задач требуется последовательно решить несколько подзадач [3]: обнаружить объект на изображении, определить его точное местонахождения, определить ключевые точки для дальнейшей классификации, классифицировать действие. Задачи обнаружение и сегментации объектов решены достаточно неплохо, с помощью специальных моделей, так называемых «региональных (областных) свёрточных нейронных сетей» (R-CNN)[4]. Для моделей типа R-CNN также предложены модификации, которые позволяет обеспечить одновременное нахождение ключевых точек изучаемого объекта на основе Mask R-CNN [5] – сеть Keypoint R-CNN [6]. Однако для обучения моделей типа Mask R-CNN требуются специальные наборы данных, в которых полностью размечены ключевые точки исследуемых объектов. Кроме того, на результат обучения, как и в любом другом случае, оказывают сильное влияние гиперпараметры модели. В данной работе предлагается метод аугментации данных, который позволяет увеличить точность нахождения ключевых точек, а также проводится сравнение различных опорных моделей с целью выявления наиболее подходящей модели для выявления ключевых точек биологического объекта.

II. Основные элементы исследования

A. *Архитектура Keypoint R-CNN*

Архитектура сети состоит из двух крупных блоков: опорной модели – свёрточной нейронной сети, которая выявляет признаки на изображении (в работе [5] для такой модели используются модификации сети ResNet) – и головной модели – свёрточной нейронной сети, которая позволяет сегментировать изображение с помощью предзаданных масок.

Основным отличием Keypoint R-CNN от Mask R-CNN является структура маски, которая используется в головной модели. Каждая маска в ней фактически представляет собой позицию одной конкретной точки на изображении, а не размеченную область расположения объекта.

Обучение модели Keypoint R-CNN осуществляется с помощью применения технологий перенесённого обучения (Transfer Learning). Опорная модель нейронной сети может меняться, при этом обучения опорной модели не происходит, а оставлявшая часть модели (головная модель) переучивается на новых данных при смене опорной модели.

B. *Набор данных для обучения*

В качестве биологического объекта в данной работе предлагается использовать животное, в качестве набора данных используется набор данных «Animal Pose» [7], в котором собран набор изображений объектов с аннотациями, в которых указаны ключевые точки объекта или область огранивающего прямоугольника, в которой на рассматриваемом изображении находится объект. Всего в наборе более 6000 объект на 4000 изображениях, однако, для задачи, решаемой в рамках данной работы, подходят только 967 изображений с пятью видами объектов: коты, коровы, собаки, лошади, овцы. Это связано с тем, что только к данным изображениям приводится аннотация, содержащая информацию о ключевых точках объекта. К каждому из выбранных изображений приводится положение следующих ключевых точек: левый глаз, правый глаз, горло, нос, холка, правая и левая ушные раковины, хвост, четыре коленных сгиба, четыре позиции стопы, четыре точки начала конечностей (все объекты в наборе – четырёхногие).

C. *Метод аугментации данных*

Чтобы увеличить количество доступных данных, предложен специальный метод аугментации данных, который основан на случайном изменении положения точки на изображении. Каждая точка изображения сдвигается относительно исходной позиции на некоторую дельту.

Предположим, что рассматривается изображение I размера $I_w = w$ пикселей по ширине и размера $I_h = h$ пикселей по ширине. На изображении отмечено n точек, у каждой из которых есть фиксированная позиция (x_i, y_i). Каждая точка обозначает опорную точку скелета объекта. Тогда на основании этих данных можно создать новый набор точек, который смещается относительно исходного на значение, определяемое произведением малого коэффициента δ на ширину и высоту изображения.

То есть

$$\forall i \in [1, n]: (x_i^{new} = x_i \pm \delta * w, y_i^{new} = y_i \pm \delta * h).$$

Таким образом, можно значительно расширить набор исходных данных, что позволит увеличить обучающую и тестовую выборки.

На рис. 1 представлен пример сетки для аугментации данных. Красным отмечена исходная точка, синей линией условное направление соединения с другими точками скелета.

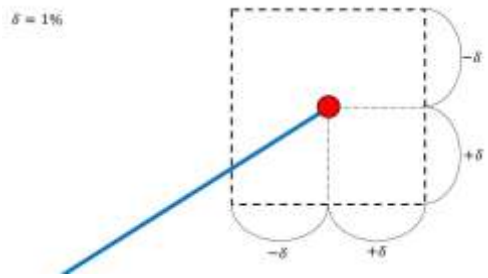

Рис. 1. Сетка для аугментации даных на примере крайней точки скелета

D. *Аугментация по классам действия*

Каждый объект в рамках рассматриваемого набора данных совершает одно из пяти действий: лежит, идёт, сидит, стоит, прыгает. Для каждого из пяти указанных действий ключевые точки объекта расположены по-разному. При этом выборка с точки зрения действий не сбалансирована: объектов, которые стоят, примерно столько же, сколько и в сумме объектов, которые совершают другие действия. Распределение объектов по классам можно видеть на рис. 2.

На качество распознавания ключевых точек объекта может влиять распределение объектов по типам действий. Исходя из этого, аугментация данных может происходить так, чтобы уравнять количество классов объектов с разным действием.

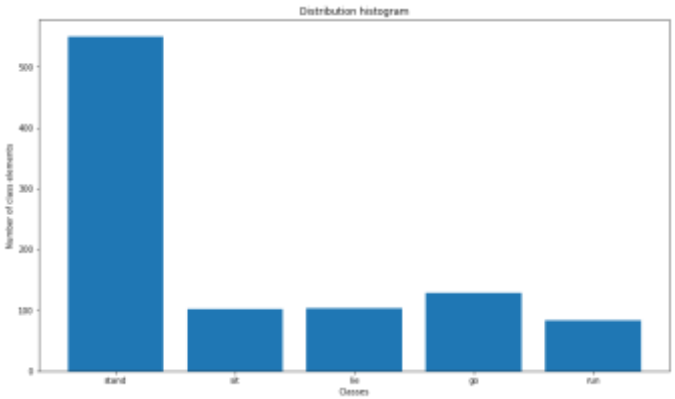

Рис. 2. Распределение объектов в наборе данных по типам совершаемых ими действий

E. *Сетка гиперпараметров для проверкеи эффективности обученных моделей*

Для проверки эффективности работы нейронной сети, распознающей точки скелета предлагается следующая сетка гиперпараметров:

1. Аугментация: без аугментации, количественная аугментация в 3 раза (для каждого из представленных изображений проводится подготовка ещё двух изображений с позициями ключевых точек, полученных с использованием метода аугментации из данной работы), аугментация по классам действий.
2. Оптимизаторы: Adam[8], SGD (стохастический градиентный спуск).
3. Скорости обучения: 0.001, с уменьшением скорости обучения (StepLR)
4. Размеры набора картинок для работы оптимизаторов (размеры батчей): 1, 3.
5. Модели: MobileNet (V2, V3 Large, V3 Small), ResNet (18, 34, 50, 101).
6. Количество эпох: 10, 15 и 20.

В качестве метрики качества была выбрана точность, которая рассчитывалась следующим образом:

предположим, что для предсказанных координат (x_i, y_i) точки i на кадре j истинным значением координат являются x_i^* и y_i^* соответственно. Тогда точность их определения рассчитывается формуле:

$$acc_i = \frac{2 - \frac{|x_i - x_i^*|}{w} - \frac{|y_i - y_i^*|}{w}}{2}.$$

Тогда точность одного кадра вычисляется как средняя точность для всех точек на кадре.

Также для проведения анализа проводился подсчёт среднего времени обработки одного кадра.

III. Основные результаты

Всего в рамках сетки гиперпараметров в данной работе было проведено 336 экспериментов. В таблице 1 представлены средние результаты среди всех экспериментов с данными гиперпараметрами и лучшие результаты работы сети при различных гиперпараметрах. Время в работе измеряется в секундах.

ТАБЛИЦА I Гиперпараметры сети и полученные результаты (точность и время работы)

Параметр сети	**Средняя точность**	**Лучшая точность**	**Среднее время**	**Лучшее время**
Без аугментации	0,8251	0,9351	0,1526	0,0350
Количественная аугментация в 3 раза	0,8355	0,9191	0,1290	0,0252
Оптимизатор Adam	0,8061	0,9191	0,1445	0.0337
Оптимизатор SGD	0,8545	0,9351	0,1370	0.0252
Размер батча 3	0,8295	0,9253	0,1475	0.0252
Размер батча 6	0,8311	0,9351	0,1340	0.0292
LR: 0,001	0,8280	0,9351	0,1590	0,0252
StepLR	0,8326	0,9253	0,1456	0,0298
Количество эпох 5	0,8240	0,9186	0,1487	0,0320
Количество эпох 10	0,8323	0,9284	0,1393	0,0274
Количество эпох 15	0,8347	0,9351	0,1344	0,0252

На основании анализа полученных результатов можно сделать следующие выводы:

1. Использование количественной аугментации данных для обучения предложенным методом повышает среднюю точность работы модели. Это соотносится с другими исследованиями о влиянии аугментации на качество обучения [9].
2. Использование SGD в процессе обучения заметно повышает точность работы моделей в среднем. Подобный результат может быть связан с общим малым количеством эпох обучения.
3. Больший размер батча повышает качество обучения нейронной сети, это соотносится с результатами других подобных работ [10].
4. Постепенное изменение шага обучение повышает точность работы нейронной сети, что тоже соотносится с другими работами [11].
5. Большее количество эпох приводит к большей точности итоговой работы нейронной сети.

Отдельно необходимо отметить, что обучение на данных, аугментированных по классам действия не приводит к заметному отличию точности работы нейронной сети от случая обучения на данных, полученных методом простой количественной аугментации. Аугментация по классам действий целесообразна в случае использования для решений задачи классификации и определения классов действий, что находится за рамками данной работы.

Среднее и лучшее время работы слабо зависит от параметров, за исключением случая использования аугментации данных.

В табл. 2 представлены лучшие результаты на всём наборе гиперпараметров, полученные при обучении сети с различными опорными моделями. Приводятся те же типы результатов работы, что и в табл. 1.

ТАБЛИЦА II Различные модели нейронных сетей, используемые в качестве опорных моделей и полученные результаты (точность и время работы).

Параметр сети	**Средняя точность**	**Лучшая точность**	**Среднее время**	**Лучшее время**
Mobile Net V2	0,6857	0,7517	0,0906	0,0421
Mobeile net V3 Large	0,8865	0,9191	0,0629	0,0371
Mobile Net V3 Small	0,8760	0,9066	0,0436	0,0252
Resnet 18	0,8502	0,9185	0,2451	0,1218
Resnet 34	0,8567	0,9253	0,1718	0,0917
Resnet 50	0,8438	0,9351	0,1617	0,1284
Resnet 101	0,8131	0,9015	0,2098	0,1333

Наилучшая средняя точность достигается сетью с опорными моделями типа MobileNet V3. Однако лучшей точности удалось достичь при использовании в качестве опорной модели Resnet 50. При этом среднее время работы на сетях MovileNet V3 в 3–4 раза меньше, чем среднее время работы с опорной моделью Resnet 50.

Между различными моделями чётко видна разница, связанная с влиянием аугментации. Так использование модели семейства ResNet не требуют дополнительных данных, и сеть достигают лучших показателей при стандартном наборе данных, тогда как при использовании опорных моделей семейства MobileNet требуется аугментация данных для получения более точных результатов.

Общее распределение моделей разного типа относительно точности и времени работы и производительности при обработке изображений представлено на рис. 3.

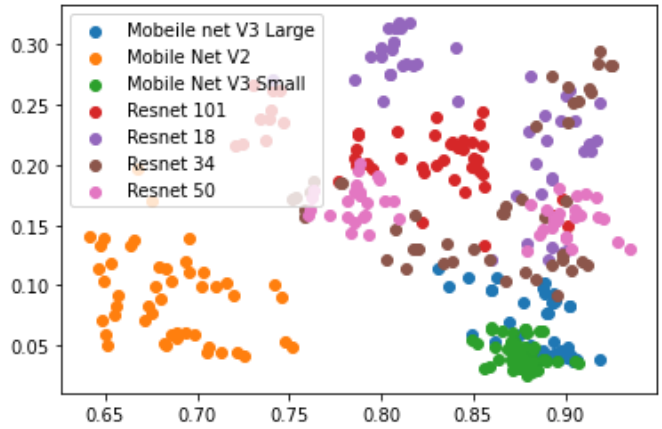

Рис. 3. Результаты экспериментов с различными моделями. По оси Y – время обработки одного кадра. По оси X – точность работы

IV. Обсуждение результатов и заключение

В работе показано, что количественная аугментация предложенным методом повышает среднюю точность обучения нейронной сети для определения ключевых точек скелета объекта. Кроме того, показано, что без перебора гиперпараметров невозможно добиться наилучшей точности и производительности работы семейства моделей Keypoint RCNN. Общая средняя точность по всем экспериментам выборки составляет 0.8303 пункта, что на 0.1 меньше, чем наилучшее значение. Таким образом, без перебора гиперпараметров невозможно добиться эффективного принятия решений в процессе управления и взаимодействия технических систем с объектам.

Лучший результат работы Keypoint R-CNN при использовании опорных моделей семейства ResNet достигается при работе с неаргументированными данными, и оптимизатором SGD, тогда как при использовании MobileNet лучший результат получается при использовании количественно-аугментированных данных и оптимизатора Adam.

Аугментация по классам действий оказалась невостребованной в данной работе. Её эффективность требует исследования в дальнейших работах, целью которых должна стать непосредственная классификация действий объекта.

В работе не получено оптимальное количество эпох обучения для достижения наибольшей точности работы сетей Keypoint R-CNN, при этом увеличение количества эпох обучения с 5 до 15 приводит к росту точности работы сетей на тестовых данных.

Используя результаты данной работы, предлагается в дальнейшем проводить классификацию действий биологических объектов для принятия управленческих решений в технических системах.

Список литературы

[1] Jin S., Jin H. Optimization of Motion Estimation Algorithm Based on FPGA Hardware System and Video Tracking // Microprocessors and Microsystems. 2021. Т. 82. С. 103867.

[2] Pereira T.D. et al. Fast animal pose estimation using deep neural networks //Nature methods. 2019. Т. 16. №. 1. С. 117-125.

[3] Nath T. et al. Using DeepLabCut for 3D markerless pose estimation across species and behaviors // Nature protocols. 2019. Т. 14. №. 7. С. 2152-2176.

[4] Ren S. et al. Faster r-cnn: Towards real-time object detection with region proposal networks //arXiv preprint arXiv:1506.01497. 2015.

[5] He K. et al. Mask r-cnn // Proceedings of the IEEE international conference on computer vision. 2017. С. 2961-2969.

[6] Ding X. et al. Local keypoint-based Faster R-CNN // Applied Intelligence. 2020. Т. 50. №. 10. С. 3007-3022.

[7] Cao J. et al. Cross-domain adaptation for animal pose estimation // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019. С. 9498-9507.

[8] Kingma D.P., Ba J. Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980. 2014.

[9] Korzhebin T.A., Egorov A.D. Comparison of Combinations of Data Augmentation Methods and Transfer Learning Strategies in Image Classification Used in Convolution Deep Neural Networks // 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus). IEEE, 2021. С. 479-482.

[10] Smith S.L. et al. Don't decay the learning rate, increase the batch size // arXiv preprint arXiv:1711.00489. 2017.

[11] Zeiler M.D. Adadelta: an adaptive learning rate method // arXiv preprint arXiv:1212.5701. 2012.